Estimating the Error Distribution in Semiparametric Transformation Models

Cédric Heuchenne
University of Liège and Université catholique de Louvain *

Rawane Samb
Université catholique de Louvain †

Ingrid Van Keilegom
Université catholique de Louvain ‡

October 8, 2015

Abstract

In this paper we consider the semiparametric transformation model $\Lambda_{\theta_o}(Y) = m(X) + \varepsilon$, where θ_o is an unknown finite dimensional parameter, the function $m(\cdot) = \mathbb{E}(\Lambda_{\theta_o}(Y)|X = \cdot)$ is "smooth" but otherwise unknown, and the covariate X is independent of the error ε. An estimator of the distribution function of ε is investigated and its weak convergence is proved. The proposed estimator depends on a profile likelihood estimator of θ_o and a nonparametric kernel estimator of m. We also evaluate the practical performance of our estimator in a simulation study for several models and sample sizes. Finally, the method is applied to a data set on the scattering of sunlight in the atmosphere.

Keywords: Empirical distribution function; Kernel smoothing; Nonparametric regression; Profile likelihood estimator; Semiparametric regression; Transformation model.

*C. Heuchenne (corresponding author: C.Heuchenne@ulg.ac.be) acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

†R. Samb acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy).

‡I. Van Keilegom acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

1 Introduction

Consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent copies of a bivariate random vector (X, Y), that satisfies the semiparametric transformation model

$$\Lambda_{\theta_o}(Y) = m(X) + \varepsilon, \tag{1.1}$$

where ε is independent of X and $\mathbb{E}(\varepsilon) = 0$. Here, $\{\Lambda_\theta : \theta \in \Theta\}$ (with $\Theta \subset \mathbb{R}^p$ compact) is a parametric family of strictly increasing functions defined on an unbounded subset $\mathcal{D}$ of $\mathbb{R}$, and m is the unknown regression function belonging to an infinite dimensional parameter set $\mathcal{M}$. We assume that $\mathcal{M}$ is a space of functions endowed with the norm $\|\cdot\|_{\mathcal{M}} = \|\cdot\|_\infty$. We denote $\theta_o \in \Theta$ and $m \in \mathcal{M}$ for the true unknown finite and infinite dimensional parameters, and we define the function $m_\theta(x) = \mathbb{E}(\Lambda_\theta(Y)|X = x)$ and the error $\varepsilon_\theta = \varepsilon(\theta) = \Lambda_\theta(Y) - m_\theta(X)$ for arbitrary $\theta \in \Theta$. Clearly, $m_{\theta_o} \equiv m$.

Our objective in this paper is to estimate the cumulative distribution function (c.d.f.) $F_\varepsilon(t) = \mathbb{P}(\varepsilon \leq t)$. Our estimation approach is based on a two-step strategy which, in a first step, replaces the unobserved regression errors ε_i's by semiparametric estimators $\widehat{\varepsilon}_i(\widehat{\theta}) = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\theta}}(X_i)$, where $\widehat{\theta}$ and $\widehat{m}_{\widehat{\theta}}$ are suitable estimators of θ_o and m_{θ_o} respectively. In a second step, the distribution function F_ε is estimated by the empirical distribution function of the $\widehat{\varepsilon}_i(\widehat{\theta})$'s as if they were the true errors. To estimate θ_o we use a profile likelihood (PL) approach, developed in Linton, Sperlich and Van Keilegom (2008), whereas for each fixed θ, $m_\theta(x)$ is estimated by means of the Nadaraya-Watson (1964) method.

To the best of our knowledge, the estimation of the distribution of the error ε in model (**??**) has not yet been investigated in the statistical literature. However, it may be very useful in various regression problems. First, taking transformations of the data may induce normality and error variance homogeneity in the transformed model. So the estimation of the error distribution in the transformed model may be used for testing these hypotheses. It may also be used for goodness-of-fit tests of a specified error distribution in a parametric or semiparametric regression setting, for testing the symmetry of the error distribution, or for various other testing problems, like tests for the parametric form of the regression or variance function, tests for comparing two regression functions, tests for the validity of the model, etc. Hence, the error distribution plays a very important role in model (**??**), both for exploratory analyses and for statistical inference.

There exists a large literature on the estimation of model (**??**) when the regression function m is parametric. A major contribution to this methodology was made by Box and Cox (1964), who proposed a parametric power family of transformations that includes the logarithm and the identity. Lots of effort has been devoted to the investigation of the Box-Cox transformation since its introduction. See, for example,

Chen, Lockhart and Stephens (2002), Freeman and Modarres (2005), Shin (2008), and Fitzenberger, Wilke and Zhang (2010) for some of the more recent references. Other dependent variable transformations have been suggested, see for example, Zellner and Revankar (1969), Manly (1976), Bickel and Doksum (1981), and the Arcsinh transformation discussed in Johnson (1949) and more recently in Robinson (1991). See also the book of Carroll and Ruppert (1988) and the review paper by Sakia (1992) for more details and references on parametric transformation models.

Over the last ten years a lot of research has been done on estimation and testing problems under model (**??**) when Λ_{θ_o} is known and equals the identity function. The starting point was the paper by Akritas and Van Keilegom (2001), who studied the estimation of the error distribution under the model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1.2}$$

i.e. a heteroscedastic version of model (**??**) with $\Lambda_{\theta_o} \equiv id$. They showed the weak convergence of their estimator of the error distribution. Their results were generalized by Neumeyer and Van Keilegom (2010) to the case where the covariate is multi-dimensional. When $\sigma(X) \equiv 1$, Müller, Schick and Wefelmeyer (2004) investigated linear functionals of the error distribution whereas the same authors estimated this distribution in partial linear models (see Müller, Schick and Wefelmeyer, 2007). The estimator of Akritas and Van Keilegom (2001) has been used in various testing problems related to model (**??**). See e.g. Neumeyer and Dette (2007), Dette et al. (2007), Pardo-Fernández et al. (2007), Cheng and Sun (2008), Dette et al. (2009), Neumeyer and Pardo-Fernández (2009), Heuchenne and Van Keilegom (2010), among many others. Tests for the validity of model (**??**) have been developed in Einmahl and Van Keilegom (2008a,b), Neumeyer (2009b) and Hlávka et al. (2011), whereas the consistency of a smooth bootstrap procedure has been shown by Neumeyer (2009a). Finally, model (**??**) has also been applied in other contexts, like e.g. for estimating ROC curves (see González-Manteiga et al., 2011) and for estimating the production frontier in efficiency analysis, where one analyzes how firms transform their inputs to produce a set of outputs (see Florens et al., 2014).

A major element of our estimation procedure is the estimation of the parameter θ_o. As mentioned before, we will make use of the results in Linton, Sperlich and Van Keilegom (2008) to this end. In the latter paper, the authors propose two estimation approaches for θ_o. The first approach is a semiparametric profile likelihood (PL) approach, whereas the second one is based on a 'mean squared distance from independence (MD)'-idea using the estimated distributions of X, ε_θ and (X, ε_θ). Linton, Sperlich and Van Keilegom (2008) derived the asymptotic distributions of their estimators under certain regularity conditions, and proved that both estimators of θ_o are asymptotically normal. The authors also showed that, in practice, the PL method

outperforms the MD method. For this reason, we focus in this paper on the PL method.

The remainder of the paper is organized as follows. In Section 2 we introduce some notations and give the precise definition of our estimator of the error distribution. In Section 3 we present the main asymptotic results of the paper, together with the assumptions under which they are valid. The results of a simulation study are given in Section 4, Section 5 is devoted to the analysis of a data set on the scattering of sunlight in the atmosphere and the proofs of the main results are collected in Section 6 and in two appendices.

2 The estimator

Our estimation procedure for the error distribution F_ε consists of two steps. In a first step, we estimate the finite dimensional parameter θ_o. This parameter is estimated by the profile likelihood (PL) method, studied in Linton, Sperlich and Van Keilegom (2008). To this end, note that under model (**??**), we have

$$\mathbb{P}\left(Y \leq y|X\right) = \mathbb{P}\left(\Lambda_{\theta_o}(Y) \leq \Lambda_{\theta_o}(y)|X\right) = \mathbb{P}\left(\varepsilon_{\theta_o} \leq \Lambda_{\theta_o}(y) - m_{\theta_o}(X)|X\right) = F_\varepsilon\left(\Lambda_\theta(y) - m_{\theta_o}(X)\right).$$

Therefore

$$f_{Y|X}(y|x) = f_\varepsilon\left(\Lambda_{\theta_o}(y) - m_{\theta_o}(x)\right)\Lambda'_{\theta_o}(y),$$

where f_ε and $f_{Y|X}$ are the densities of ε, and of Y given X, respectively. Then, the log likelihood function with respect to $\theta \in \Theta$ is given by

$$\sum_{i=1}^{n}\left\{\log f_{\varepsilon_\theta}(\Lambda_\theta(Y_i) - m_\theta(X_i)) + \log\Lambda'_\theta(Y_i)\right\}, \tag{2.1}$$

where f_{ε_θ} is the density function of ε_θ. The idea of the PL method is to replace all unknown expressions in the likelihood function by nonparametric kernel estimators. For this, let

$$\widehat{m}_\theta(x) = \frac{\sum_{i=1}^{n}\Lambda_\theta(Y_i)K_1\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^{n}K_1\left(\frac{X_i-x}{h}\right)} \tag{2.2}$$

be the Nadaraya-Watson (1964) estimator of $m_\theta(x)$ based on the 'responses' $\Lambda_\theta(Y_i)$, $i = 1, \ldots, n$, and let

$$\widehat{f}_{\varepsilon_\theta}(t) = \frac{1}{ng}\sum_{i=1}^{n}K_2\left(\frac{\widehat{\varepsilon}_i(\theta) - t}{g}\right) \tag{2.3}$$

be a kernel estimator of the density of $\varepsilon(\theta)$, where $\widehat{\varepsilon}_i(\theta) = \Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i)$. Here, K_1 and K_2 are kernel functions and h and g are appropriate bandwidth sequences, tending to zero as n tends to infinity. This leads to the following PL estimator of θ_o:

$$\widehat{\theta} = \arg\max_{\theta\in\Theta}\sum_{i=1}^{n}\left[\log\widehat{f}_{\varepsilon_\theta}(\Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i)) + \log\Lambda'_\theta(Y_i)\right]. \tag{2.4}$$

Since the estimator $\widehat{m}_\theta(X_i)$ converges to $m_\theta(X_i)$ at a slower rate for those X_i that are close to the boundary of the support $\mathcal{X}$ of X, we assume implicitly that the estimator $\widehat{\theta}$ trims the observations X_i that are outside a subset $\mathcal{X}_0$ of $\mathcal{X}$. Note that by doing so, we keep the root-n consistency of $\widehat{\theta}$ proved in Linton, Sperlich and Van Keilegom (2008).

Next, we use the estimator $\widehat{\theta}$ to build the estimated residuals $\widehat{\varepsilon}_i(\widehat{\theta}) = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\theta}}(X_i)$ (where as above, observations X_i that are outside $\mathcal{X}_0$ are not considered). Then, our proposed estimator $\widehat{F}_{\widehat{\varepsilon}}(t)$ for $F_\varepsilon(t)$ is defined by

$$\widehat{F}_{\widehat{\varepsilon}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(\widehat{\varepsilon}_i(\widehat{\theta}) \leq t\right). \tag{2.5}$$

In order to obtain the asymptotic distribution of this estimator, we will also need the (unfeasible) estimator $\widehat{F}_\varepsilon(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\varepsilon_i \leq t)$, based on the true, but unknown errors $\varepsilon_i = \varepsilon_i(\theta_o) = \Lambda_{\theta_o}(Y_i) - m(X_i)$. It will turn out that both the expressions $\widehat{F}_{\widehat{\varepsilon}}(t) - \widehat{F}_\varepsilon(t)$ and $\widehat{F}_\varepsilon(t) - F_\varepsilon(t)$ contribute to the asymptotic distribution of the estimator $\widehat{F}_{\widehat{\varepsilon}}(t)$ (see Section 3 for more details).

Remark 2.1

The estimator $\widehat{F}_{\widehat{\varepsilon}}(\cdot)$ in (**??**) could be compared with a classical integrated density estimator of the type (**??**), where θ is replaced by $\widehat{\theta}$ in (**??**). This idea has already been studied in a number of other contexts (see e.g. Reiss, 1981, among others). It is expected that this alternative estimator of the error distribution has the same asymptotic distribution as the original estimator $\widehat{F}_{\widehat{\varepsilon}}(\cdot)$, provided the new bandwidth coming from the final kernel density estimator (which invokes an additional bias compared to the original estimator) is chosen in an appropriate way. In practice, the alternative estimator will have the advantage of being smooth, but on the other hand it has the important drawback that the additional bandwidth, which controls the smoothing of the final kernel estimator, needs to be chosen in an appropriate way in order to avoid bias effects. For the latter reason, we have preferred to focus on the original estimator $\widehat{F}_{\widehat{\varepsilon}}(\cdot)$ in this paper.

Remark 2.2

The above methodology could be extended to the multivariate case. The procedure could indeed be used with a multivariate kernel estimator instead of the estimator $\widehat{m}_\theta(x)$ in (**??**). Theoretically, the asymptotic properties of the estimator of the error distribution given in Neumeyer and Van Keilegom (2010) should be extended to the case where $\Lambda_{\theta_0}(\cdot)$ is not the identity function.

Remark 2.3

Recently, new maximization procedures with respect to both the parametric and the nonparametric com-

ponents have been developed (see for example Ding and Nan, 2011). These have been shown to improve efficiency of the resulting estimators but might suffer from numerical problems. More precisely, maximizing (with respect to 'bundled' parameters) can be achieved in a sieve space where unknown functions can be approximated by B-splines. In our log likelihood (**??**), there would be two nuisance parameters ζ_1 and ζ_2 corresponding to the unknown functions m_θ and f_{ε_θ}, that should be approximated by B-splines. Similarly to the notations used in Ding and Nan (2011), we can rewrite (**??**) as

$$\sum_{i=1}^{n} m_l(Y_i, \theta, \zeta_2(Y_i, \theta, \zeta_1(X_i, \theta))),$$

where m_l corresponds to the likelihood function for a single data point. First, the assumptions of Theorem 2.1 of Ding and Nan (2011) should be studied in the present context. Next, as it can be seen from the above expression, ζ_2 depends on ζ_1 and maximization should be achieved with respect to two nested sets of coefficients of the B-spline basis functions. It is unclear whether this will lead to a numerically stable maximization problem.

In addition, our objective is to estimate the cumulative distribution function of the errors (not the density) and therefore the residuals themselves are of primary interest. We fear that if we try to estimate everything at once, then the errors, which are the building blocks of our procedure, might be badly influenced by numerical instability problems coming from the important number of parameters that need to be estimated at once, and the whole estimation procedure might become unstable. For these reasons and since in any case, a practical automatic choice of the knots (or possibly a smoothing splines technique) has to be conducted, we finally decided to develop the strategy proposed above.

3 Asymptotic results

Before we give the main asymptotic results of this paper, we first need to introduce a number of notations, and we also give the assumptions under which these results are valid.

3.1 Notations

We denote $\mathscr{X}_n = \{(X_j, Y_j) : j = 1, \ldots, n\}$ and $F_{Y|X}(y|x) = \mathbb{P}(Y \leq y | X = x)$. When there is no ambiguity possible, we use the abbreviated notations ε and m to indicate ε_{θ_o} and m_{θ_o}. Throughout the paper, $\mathcal{N}(\theta_o)$ represents a neighborhood of θ_o. For the kernel K_j $(j = 1, 2)$ and for any q, let $\mu(q, K_j) = \int v^q K_j(v) dv$ and let $K_j^{(q)}$ be the qth derivative of K_j. For any function $\varphi_\theta(y)$, denote $\dot{\varphi}_\theta(y) =$

$\partial\varphi_\theta(y)/\partial\theta = (\partial\varphi_\theta(y)/\partial\theta_1, \ldots, \partial\varphi_\theta(y)/\partial\theta_p)^t$ and $\varphi'_\theta(y) = \partial\varphi_\theta(y)/\partial y$. Also, let $\|A\| = (A^tA)^{1/2}$ be the Euclidean norm of any vector A. For any functions $\widetilde{m}$, r, f, φ and q, and any $\theta \in \Theta$, let $s = (\widetilde{m}, r, f, \varphi, q)$, $s_\theta = (m_\theta, \dot{m}_\theta, f_{\varepsilon_\theta}, f'_{\varepsilon_\theta}, \dot{f}_{\varepsilon_\theta})$, $\varepsilon_i(\theta, \widetilde{m}) = \Lambda_\theta(Y_i) - \widetilde{m}(X_i)$, and define

$$G_n(\theta, s) = n^{-1}\sum_{i=1}^{n}\left\{\frac{1}{f\{\varepsilon_i(\theta,\widetilde{m})\}}\left[\varphi\{\varepsilon_i(\theta,\widetilde{m})\}\{\dot{\Lambda}_\theta(Y_i) - r(X_i)\} + q\{\varepsilon_i(\theta,\widetilde{m})\}\right] + \frac{\dot{\Lambda}'_\theta(Y_i)}{\Lambda'_\theta(Y_i)}\right\},$$

$G(\theta, s) = \mathbb{E}[G_n(\theta, s)]$ and $\mathcal{G}(\theta_o, s_{\theta_o}) = \frac{\partial}{\partial\theta}G(\theta, s_\theta)\big|_{\theta=\theta_o}$.

For any compact subset I in $\mathbb{R}$ with nonempty interior and for any $\alpha > 0$ and $0 < M < \infty$, let $C_M^{1+\alpha}(I)$ represent the class of all differentiable functions d defined on I such that $\|d\|_{1+\alpha} \le M$, where

$$\|d\|_{1+\alpha} = \max\left\{\sup_x |d(x)|, \sup_x |d'(x)|\right\} + \sup_{x,x'}\frac{|d'(x) - d'(x')|}{|x - x'|^\alpha},$$

and where all suprema are taken over I.

3.2 Technical assumptions

(**A1**) The function K_j $(j = 1, 2)$ is symmetric, has compact support, $\int v^k K_j(v)dv = 0$ for $k = 1, \ldots, q_j - 1$ and $\int v^{q_j} K_j(v)dv \neq 0$ for some $q_j \ge 4$, and K_j is twice continuously differentiable.

(**A2**) The bandwidths h and g satisfy $nh^{2q_1} = o(1)$, $ng^{2q_2} = o(1)$, $nh^{q_1+1}(\log h^{-1})^{-1} \to \infty$ and $ng^6(\log g^{-1})^{-2} \to \infty$ when $n \to \infty$ (where q_1 and q_2 are defined in (A1)).

(**A3**) (i) The support $\mathcal{X}$ of the covariate X is a compact subset of $\mathbb{R}$, and $\mathcal{X}_0$ is a compact subset with nonempty interior inside the interior of $\mathcal{X}$.
(ii) The density f_X is bounded away from zero and infinity on $\mathcal{X}$, and is $q_1 - 1$ times continuously differentiable.

(**A4**) The function $m_\theta(x)$ is continuously differentiable with respect to θ on $\mathcal{X} \times \mathcal{N}(\theta_0)$, and the functions $m_\theta(x)$ and $\dot{m}_\theta(x)$ are q_1 times continuously differentiable with respect to x on $\mathcal{X} \times \mathcal{N}(\theta_0)$. All derivatives are bounded, uniformly in $(x, \theta) \in \mathcal{X} \times \mathcal{N}(\theta_o)$.

(**A5**) The error $\varepsilon = \Lambda_{\theta_o}(Y) - m(X)$ has finite fourth moment, ε is independent of X and $f_\varepsilon(y) > 0$ for all y.

(**A6**) The distribution $F_{\varepsilon_\theta|X}(t|x)$ of ε_θ is three times continuously differentiable with respect to t and θ, and

$$\sup_{\theta,t,x}\left|\frac{\partial^{k+\ell}}{\partial t^k \partial\theta_1^{\ell_1}\ldots\partial\theta_p^{\ell_p}}F_{\varepsilon_\theta|X}(t|x)\right| < \infty$$

for all k and ℓ such that $0 \leq k+\ell \leq 2$, where $\ell = \ell_1 + \ldots + \ell_p$ and $\theta = (\theta_1, \ldots, \theta_p)^t$.

(**A7**) (i) The transformation $\Lambda_\theta(y)$ is three times continuously differentiable with respect to both θ and y, and there exists $\alpha > 0$ such that

$$\mathbb{E}\left[\sup_{\theta':\|\theta'-\theta\|\leq\alpha}\left|\frac{\partial^{k+\ell}}{\partial y^k \partial\theta_1^{\ell_1}\ldots\partial\theta_p^{\ell_p}}\Lambda_{\theta'}(Y)\right|\right] < \infty$$

for all $\theta \in \Theta$, and for all k and ℓ such that $0 \leq k+\ell \leq 3$, where $\ell = \ell_1 + \ldots + \ell_p$ and $\theta = (\theta_1, \ldots, \theta_p)^t$. Moreover, $\sup_{x\in\mathcal{X}} \|\mathbb{E}[\dot{\Lambda}^4_{\theta_o}(Y)|X=x]\| < \infty$.

(ii) The density function of $(\dot{\Lambda}_\theta(Y), X)$ exists and is continuous for all $\theta \in \Theta$.

(**A8**) For all $\eta > 0$, there exists $\epsilon(\eta) > 0$ such that

$$\inf_{\|\theta-\theta_o\|>\eta} \|G(\theta, s_\theta)\| \geq \epsilon(\eta) > 0.$$

Moreover, the matrix $\mathcal{G}(\theta_o, s_{\theta_o})$ is non-singular.

(**A9**) $\mathbb{E}(\Lambda_{\theta_o}(Y)) = 1$, $\Lambda_{\theta_o}(0) = 0$ and the set $\{x \in \mathcal{X}_0 : m'(x) \neq 0\}$ has nonempty interior.

Assumptions (A1), part of (A2), (A3) (i) and part of (A3)(ii), (A4), (A6), part of (A7)(i) and (A8) are used by Linton, Sperlich and Van Keilegom (2008) to show that the PL estimator $\widehat{\theta}$ of θ_o is root n-consistent. Part of assumptions (A2), (A3) (ii) and (A7) (i), assumptions (A5) and (A7) (ii), are needed to obtain the uniform convergence rates of the Nadaraya-Watson estimator $\widehat{m}_{\widehat{\theta}}(x)$ and its derivatives with respect to x and θ. Finally, (A9) is needed for identifying the model (see Vanhems and Van Keilegom (2013)).

3.3 Main results

The estimator $\widehat{F}_{\widehat{\varepsilon}}(t)$ is not a sum of independent terms. Therefore, we start by constructing an asymptotic representation for $\widehat{F}_{\widehat{\varepsilon}}(t)$, which decomposes $\widehat{F}_{\widehat{\varepsilon}}(t)$ in essentially four parts. The first one equals the empirical distribution function based on the true errors ε_i's, the second and third parts account for the replacement of the unknown $m_{\theta_o}(X_i)$ and $\Lambda_{\theta_o}(Y_i)$ in ε_i by $\widehat{m}_{\widehat{\theta}}(X_i)$ and $\Lambda_{\widehat{\theta}}(Y_i)$, while the last part is asymptotically negligible.

Theorem 3.1. *Assume* (A1)-(A9). *Then,*

$$\widehat{F}_{\widehat{\varepsilon}}(t) - F_\varepsilon(t) \quad = \quad n^{-1}\sum_{i=1}^{n} \phi_{\theta_o}(t, X_i, Y_i) + R_n(t),$$

where $\sup\{|R_n(t)| : -\infty < t < +\infty\} = o_{\mathbb{P}}(n^{-1/2})$,

$$\phi_{\theta_o}(t,x,y) = \mathbb{1}_{(\infty,t]}(\Lambda_{\theta_o}(y) - m(x)) - F_{\varepsilon}(t) + f_{\varepsilon}(t)(\Lambda_{\theta_o}(y) - m(x)) + \rho^t_{\theta_o}(x,y)h(t),$$

$\mathbb{1}_A(\cdot)$ *denotes the indicator function,* $\widehat{\theta} - \theta_o = \frac{1}{n}\sum_{i=1}^n \rho_{\theta_o}(X_i, Y_i) + o_{\mathbb{P}}(n^{1/2})$ *is the i.i.d. representation given in Theorem 4.1 of Linton, Sperlich and Van Keilegom (2008),* ρ^t *denotes the transpose of* ρ *and*

$$h(t) = \mathbb{E}\left[\frac{\partial}{\partial\theta} F_{\varepsilon_\theta|X}(t|X)\big|_{\theta=\theta_o}\right].$$

We continue with the statement of the weak convergence of the process $n^{1/2}(\widehat{F}_{\widehat{\varepsilon}}(t) - F_{\varepsilon}(t))$ $(-\infty < t < +\infty)$.

Corollary 3.1. *Suppose that the assumptions of* Theorem **??** *are satisfied. Then, the process* $\widehat{Z}_n(t) = n^{1/2}[\widehat{F}_{\widehat{\varepsilon}}(t) - F_{\varepsilon}(t)]$, $-\infty < t < +\infty$, *converges weakly to a zero-mean Gaussian process* $Z(t)$ *with covariance function*

$$\mathrm{Cov}\left(Z(t), Z(t')\right) = \mathbb{E}\left(\phi_{\theta_o}(t,X,Y)\phi_{\theta_o}(t',X,Y)\right).$$

Remark 3.1. It is possible to relax assumption $(A3)$ by allowing the bounds of a compact subset of the support of X to tend to infinity with n. Let c_n correspond to these bounds (according to the notations of Hansen, 2008, who treated this problem for classical kernel estimators with dependent data) and assume that $\delta_n = \inf_{|x|\leq c_n} f_X(x) > 0$ (that tends to 0 when n tends to infinity). It is then possible to show that the uniform consistency rate of Nadaraya-Watson type estimators of the regression function is multiplied by δ_n^{-1} (with respect to the case where the support of X is compact). If δ_n tends to zero sufficiently slowly, the final rates and the weak convergence of our estimator of the error distribution would then be preserved. However, the proof of this result would require two important steps. First, the asymptotic normality of the estimator $\widehat{\theta}$ of Linton, Sperlich and Van Keilegom (2008) should be extended to the case where the likelihood function only uses factors corresponding to X_i, $i = 1, \ldots, n$, included in the above subset depending on n. Second, this extension should also be achieved for our estimator of the error distribution, first for the asymptotic equicontinuity property of Lemma 1 in Appendix B and next, for the treatment of the asymptotic representation, where each term depends on n.

4 Simulations

In this section, the finite sample performance of our estimator $\widehat{F}_{\widehat{\varepsilon}}(t)$ is investigated. This is achieved through simulations described hereunder. Consider the following transformation model:

$$\Lambda_{\theta_o}(Y) = b_0 + b_1 X^2 + b_2 \sin(\pi X) + \sigma_e e, \tag{4.1}$$

where Λ_θ is the Box-Cox (1964) transformation:

$$\Lambda_\theta(y) = \begin{cases} \frac{y^\theta - 1}{\theta}, & \theta \neq 0, \\ \log(y), & \theta = 0, \end{cases}$$

X is uniformly distributed on the interval $[-1, 1]$ and $\varepsilon = \sigma_e e$ is independent of X. We choose $\theta_o = 0$, 0.5 or 1, and for each θ_0, three different models are considered:

Model 1: $b_0 = 6.5$, $b_1 = 5$, $\sigma_e = 1.5$;
Model 2: $b_0 = 4.5$, $b_1 = 3.5$, $\sigma_e = 1$;
Model 3: $b_0 = 2.5$, $b_1 = 2.5$, $\sigma_e = 0.5$.

For each model, $b_2 = b_0 - 3\sigma_e$. We study two simulation settings that correspond to two different distributions of e in (**??**): first, a standard normal distribution and, second, a mixture of the normal distributions $N(-1.5, 0.25)$ and $N(1.5, 0.25)$ with equal weights. In order to avoid negative values of $\Lambda_{\theta_o}(Y_i)$, $i = 1, \ldots, n$, these distributions are truncated on $[-3, 3]$ (namely, the corresponding densities are put to zero outside the interval $[-3, 3]$ but their integrals on this support are equal to one).

One hundred samples of sizes $n = 100$ and $n = 200$ are generated and the Epanechnikov kernel $K(x) = \frac{15}{16}(1 - x^2)^2 \mathbb{1}(|x| \leq 1)$ is used for both the estimators of the regression and the density functions. For the estimation of θ_0 and $F_\varepsilon(t)$, we proceed as follows. Let

$$L_\theta(h, g) = \sum_{i=1}^{n} \left[\log \widehat{f}_{\varepsilon_\theta}(\widehat{\varepsilon}_i(\theta, h)) + \log \Lambda'_\theta(Y_i) \right],$$

where $\widehat{\varepsilon}_i(\theta, h) = \Lambda_\theta(Y_i) - \widehat{m}_\theta(X_i, h)$ and $\widehat{m}_\theta(x, h)$ denotes $\widehat{m}_\theta(x)$ constructed with bandwidth h. This function will be maximized with respect to θ for given (optimal) values of (h, g). For each value of θ, $h^*(\theta)$ is obtained by least squares cross-validation,

$$h^*(\theta) = \arg\min_h \sum_{i=1}^{n} (\Lambda_\theta(Y_i) - \widehat{m}_{-i,\theta}(X_i))^2,$$

where

$$\widehat{m}_{-i,\theta}(X_i) = \frac{\sum_{j=1, j\neq i}^{n} \Lambda_\theta(Y_j) K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j=1, j\neq i}^{n} K\left(\frac{X_j - X_i}{h}\right)}$$

and g can be chosen with a classical bandwidth selection rule for kernel density estimation. Here, for simplicity, the normal rule is used ($\widehat{g}(\theta) = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_{\widehat{\varepsilon}(\theta,h^*(\theta))}$, where $\widehat{\sigma}_{\widehat{\varepsilon}(\theta,h^*(\theta))}$ is the classical empirical estimator of the standard deviation based on $\widehat{\varepsilon}_i(\theta, h^*(\theta))$, $i = 1, \ldots, n$). The solution

$$\widehat{\theta} = \arg\max_{\theta} L_\theta(h^*(\theta), \widehat{g}(\theta))$$

is therefore obtained iteratively (maximization problems are solved with the function 'optimize' in R with $h \in [0, 2]$ and $\theta \in [-20, 20]$) and the estimator of $F_\varepsilon(t)$ is finally given by

$$\widehat{F}_{\widehat{\varepsilon}}(t) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(\widehat{\varepsilon}_i(\widehat{\theta}, h^*(\widehat{\theta})) \leq t\right).$$

Figure **??** shows realizations of the estimator $\widehat{F}_{\widetilde{e}}(t)$ (the above empirical estimator but based on standardized residuals $\widetilde{e}_i = \widehat{\varepsilon}_i(\widehat{\theta}, h^*(\widehat{\theta}))/\sigma_e$, instead of $\widehat{\varepsilon}_i(\widehat{\theta}, h^*(\widehat{\theta}))$, $i = 1, \ldots, n$) when the error distribution is normal and when it is a mixture of two normals. Tables **??**, **??** and **??** show the bias, the variance (Var) and the mean squared error (MSE) of $\widehat{F}_{\widetilde{e}}(t)$ for these two error distributions and for a number of values of t. Dividing the residuals by σ_e only aims at comparing models 1, 2 and 3 (modes are the same); in practice, if we would rather construct a standardized version of $\widehat{F}_{\widehat{\varepsilon}}(t)$, a (global) estimator of σ_e should be introduced in the procedure (see Section 5). Moreover, Tables **??** and **??** show the integrated mean squared error (IMSE) of $\widehat{F}_{\widetilde{e}}(t)$ for both assumed error distributions.

As expected, we can observe (in particular from Tables **??** and **??**) that estimation improves for sample sizes going from $n = 100$ to $n = 200$ and is better for the normal error density than for the mixture. These tables also suggest that a larger σ_e globally leads to worse results. In these simulated examples, the best results are obtained for the logarithmic transformation. This is intuitively clear, because the shape of the logarithmic function is very different from a power function (the range of $\log(y)$ equals $(-\infty, +\infty)$ and $\log(y)$ is very steep close to $y = 0$, while $(y^\theta - 1)/\theta$ takes values from $-1/\theta$ to $+\infty$ for a given $\theta > 0$ and is less steep close to $y = 0$). Therefore, if the logarithm is the true transformation, then it should be relatively easy to detect. This is in line with the findings in Linton, Sperlich and Van Keilegom (2008), who reported the MSE of $\widehat{\theta}$ for several values of θ.

5 Data analysis

We apply our testing procedure to a data set composed of 355 observations resulting from an experiment on the scattering of sunlight in the atmosphere (see Bellver, 1987). The data can be found in Cleveland (1993).

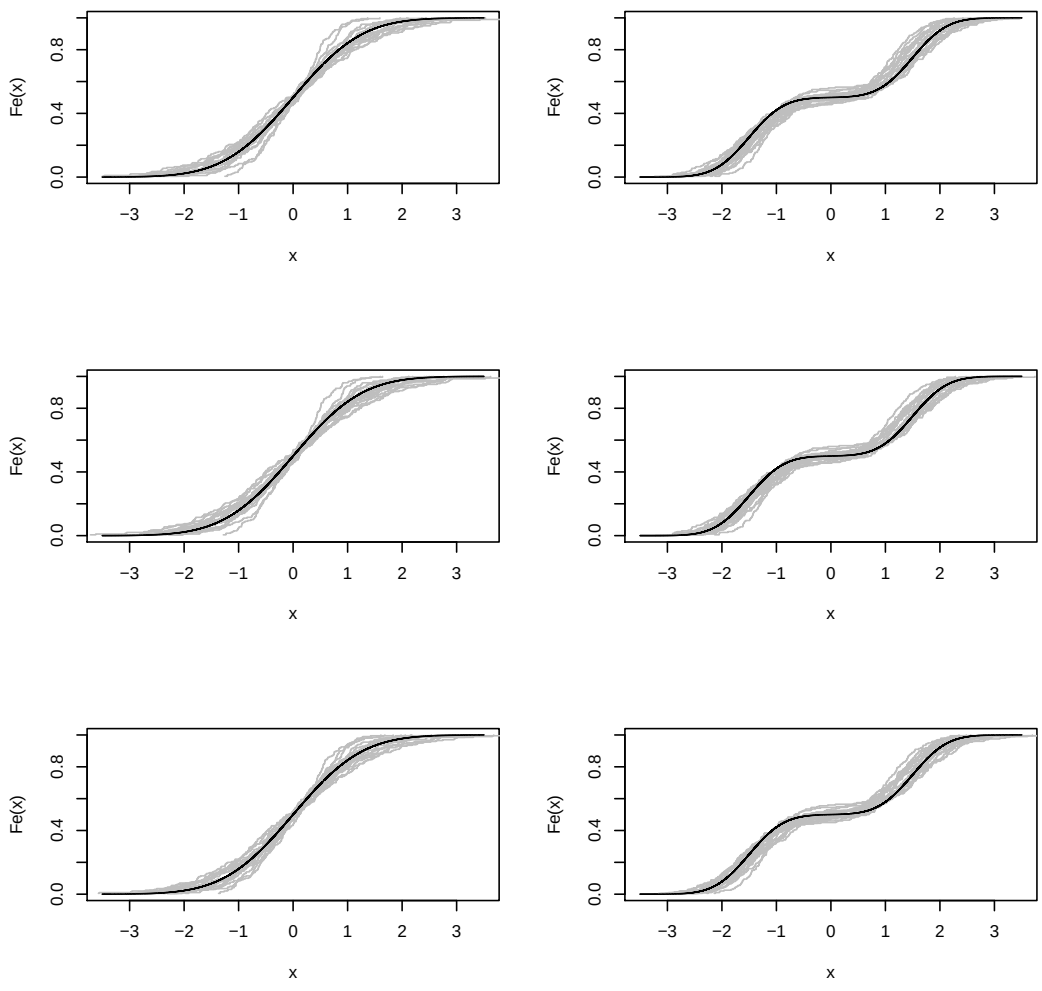

Figure 1: Realizations of $\widehat{F}_{\widetilde{e}}(t)$ for $n = 200$ and $\theta_0 = 0$, when the distribution of e is either a standard normal (on the left) or a mixture of two normal distributions ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights (on the right). The first row corresponds to model 1, the second row to model 2 and the third row to model 3.

Model	θ_0		$n = 100$			$n = 200$		
			$\widehat{F}_{\widetilde{e}}(-1)$	$\widehat{F}_{\widetilde{e}}(0)$	$\widehat{F}_{\widetilde{e}}(1)$	$\widehat{F}_{\widetilde{e}}(-1)$	$\widehat{F}_{\widetilde{e}}(0)$	$\widehat{F}_{\widetilde{e}}(1)$
		Bias	-2.13	-0.74	2.33	-0.63	-0.68	0.66
	$\theta_0 = 0$	Var	54.86	10.74	60.90	35.09	4.18	34.56
		MSE	59.38	11.29	66.31	35.49	4.63	34.99
$b_0 = 6.5$		Bias	-2.20	-0.88	2.57	-0.40	-0.61	0.48
$b_1 = 5$	$\theta_0 = 0.5$	Var	128.71	10.73	126.17	75.15	4.22	77.49
$\sigma_e = 1.5$		MSE	133.53	11.51	132.75	75.31	4.59	77.73
		Bias	-2.38	-0.87	2.90	-0.11	-0.59	0.31
	$\theta_0 = 1$	Var	152.11	10.98	146.01	101.49	4.65	96.59
		MSE	157.75	11.74	154.39	101.50	5.00	96.68
		Bias	-2.25	-0.74	2.34	-0.78	-0.64	0.62
	$\theta_0 = 0$	Var	49.29	11.27	52.96	27.26	4.31	30.33
		MSE	54.33	11.81	58.41	27.87	4.71	30.72
$b_0 = 4.5$		Bias	-1.86	-0.75	2.07	-0.47	-0.68	0.54
$b_1 = 3.5$	$\theta_0 = 0.5$	Var	112.15	11.60	107.15	62.40	4.14	61.93
$\sigma_e = 1$		MSE	115.59	12.17	111.42	62.62	4.59	62.21
		Bias	-1.54	-0.76	2.08	-0.64	-0.64	0.65
	$\theta_0 = 1$	Var	139.96	11.42	135.58	88.93	4.29	85.20
		MSE	142.35	11.99	139.89	89.34	4.70	85.62
		Bias	-1.48	-0.52	1.29	-0.96	-0.69	0.71
	$\theta_0 = 0$	Var	40.46	10.98	41.96	21.11	3.85	22.93
		MSE	42.64	11.25	43.62	22.02	4.32	23.43
$b_0 = 2.5$		Bias	-1.55	-0.46	1.14	-0.91	-0.72	0.65
$b_1 = 2.5$	$\theta_0 = 0.5$	Var	78.60	11.44	92.06	45.07	3.86	46.43
$\sigma_e = 0.5$		MSE	80.99	11.65	93.35	45.89	4.37	46.85
		Bias	-1.17	-0.58	1.25	-0.78	-0.76	0.55
	$\theta_0 = 1$	Var	100.62	11.70	103.77	56.34	3.81	56.16
		MSE	101.98	12.04	105.33	56.95	4.38	56.46

Table 1: Bias($\widehat{F}_{\widetilde{e}}(t)$) ($\times 10^2$), Var($\widehat{F}_{\widetilde{e}}(t)$) ($\times 10^4$) and $MSE(\widehat{F}_{\widetilde{e}}(t))$ ($\times 10^4$) for different models, values of t and sample sizes, when e has a standard normal density.

Model	θ_0	$n = 100$	$n = 200$
$b_0 = 6.5$	$\theta_0 = 0$	28.20	15.78
$b_1 = 5$	$\theta_0 = 0.5$	82.73	36.80
$\sigma_e = 1.5$	$\theta_0 = 1$	95.59	56.26
$b_0 = 4.5$	$\theta_0 = 0$	25.13	13.04
$b_1 = 3.5$	$\theta_0 = 0.5$	59.91	30.20
$\sigma_e = 1$	$\theta_0 = 1$	80.95	49.67
$b_0 = 2.5$	$\theta_0 = 0$	20.54	10.25
$b_1 = 2.5$	$\theta_0 = 0.5$	43.77	20.59
$\sigma_e = 0.5$	$\theta_0 = 1$	56.73	26.03

Table 2: $IMSE(\widehat{F}_{\widehat{\varepsilon}}(t))$ ($\times 10^4$) for different models, values of t and sample sizes, when e has a standard normal density.

The response Y is the scattering angle at which the polarization of sunlight vanishes, called the Babinet point. Note that the response is positive, which justifies the use of a Box-Cox transformation. Moreover, the covariate X is the cube root of a measure of particulate concentration in the atmosphere and we standardize it.

This data set has already been analyzed, but without transformation of the response variable, in different articles, like in Hart (1997), in Zhang (2003) and in Van Keilegom, González-Manteiga and Sánchez-Sellero (2008). A test for linearity of the underlying regression function was realized in Hart (1997), while different tests for lth degree polynomial regression ($l = 1, 2, 3, 4$) were realized in Zhang (2003) and in Van Keilegom, González-Manteiga and Sánchez-Sellero (2008), both with their own testing procedure. Here, similarly to Section 4 and for graphical representation purposes, we compute the error distribution based on standardized residuals $\widehat{\varepsilon}_i(\widehat{\theta}, h^*(\widehat{\theta}))/\widehat{\sigma}_{\widehat{\varepsilon}(\widehat{\theta},h^*(\widehat{\theta}))}$, where $\widehat{\sigma}_{\widehat{\varepsilon}(\widehat{\theta},h^*(\widehat{\theta}))}$ is defined as the classical empirical estimator of the standard deviation based on $\widehat{\varepsilon}_i(\widehat{\theta}, h^*(\widehat{\theta}))$, $i = 1, \ldots, n$. We obtain $\widehat{\theta} = 1,9$. As we can see from Figure **??**, this estimated error distribution seems classical (close to the standard normal distribution), suggesting it makes sense to consider the applied Box-Cox transformation.

Model	θ_0		$n = 100$				
			$\widehat{F}_{\widetilde{e}}(-1.5)$	$\widehat{F}_{\widetilde{e}}(-1)$	$\widehat{F}_{\widetilde{e}}(0)$	$\widehat{F}_{\widetilde{e}}(1)$	$\widehat{F}_{\widetilde{e}}(1.5)$
		Bias	-4.83	-7.14	-0.09	5.45	3.84
	$\theta_0 = 0$	Var	57.44	27.40	20.95	25.03	49.03
		MSE	80.70	78.34	20.96	54.70	63.77
$b_0 = 6.5$		Bias	-8.33	-11.09	-0.14	9.43	7.85
$b_1 = 5$	$\theta_0 = 0.5$	Var	106.59	83.41	20.55	87.30	101.71
$\sigma_e = 1.5$		MSE	175.98	206.34	20.57	176.18	163.33
		Bias	-9.03	-12.37	-0.09	10.37	8.30
	$\theta_0 = 1$	Var	128.39	103.06	20.10	115.22	138.46
		MSE	209.93	256.01	20.11	222.70	207.35
		Bias	-4.92	-7.45	-0.16	5.78	4.00
	$\theta_0 = 0$	Var	52.07	33.71	20.04	30.88	46.08
		MSE	76.28	89.17	20.06	64.25	62.08
$b_0 = 4.5$		Bias	-7.29	-9.96	-0.18	8.38	6.85
$b_1 = 3.5$	$\theta_0 = 0.5$	Var	90.41	67.51	20.84	68.86	81.26
$\sigma_e = 1$		MSE	143.55	166.66	20.87	139.04	128.18
		Bias	-8.32	-11.09	-0.18	9.56	7.73
	$\theta_0 = 1$	Var	109.51	85.19	20.67	91.99	102.16
		MSE	178.74	208.12	20.71	183.33	161.91
		Bias	-5.73	-8.30	-0.22	7.00	5.39
	$\theta_0 = 0$	Var	54.40	34.58	21.43	36.19	49.72
		MSE	87.23	103.43	21.47	85.15	78.77
$b_0 = 2.5$		Bias	-6.81	-9.63	-0.30	8.14	6.29
$b_1 = 2.5$	$\theta_0 = 0.5$	Var	82.54	59.52	21.04	53.56	75.14
$\sigma_e = 0.5$		MSE	128.92	152.21	21.13	119.78	114.70
		Bias	-7.71	-10.80	-0.25	9.35	7.31
	$\theta_0 = 1$	Var	99.68	73.63	21.32	72.65	96.30
		MSE	159.13	190.21	21.38	160.02	149.73

Table 3: Bias($\widehat{F}_{\widetilde{e}}(t)$) ($\times 10^2$), Var($\widehat{F}_{\widetilde{e}}(t)$) ($\times 10^4$) and $MSE(\widehat{F}_{\widetilde{e}}(t))$ ($\times 10^4$) for different models, values of t and $n = 100$, when the distribution of e is a mixture of two normal densities ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights.

Model	θ_0		$n = 200$				
			$\widehat{F}_{\widetilde{e}}(-1.5)$	$\widehat{F}_{\widetilde{e}}(-1)$	$\widehat{F}_{\widetilde{e}}(0)$	$\widehat{F}_{\widetilde{e}}(1)$	$\widehat{F}_{\widetilde{e}}(1.5)$
		Bias	-2.47	-3.93	0.01	3.22	1.82
	$\theta_0 = 0$	Var	37.46	12.96	10.34	11.20	28.77
		MSE	43.53	28.42	10.34	21.58	32.06
$b_0 = 6.5$		Bias	-4.69	-5.88	-0.02	4.72	4.43
$b_1 = 5$	$\theta_0 = 0.5$	Var	81.25	33.96	10.38	32.59	73.38
$\sigma_e = 1.5$		MSE	103.20	68.51	10.38	54.89	93.01
		Bias	-5.57	-6.80	-0.04	5.45	5.45
	$\theta_0 = 1$	Var	105.76	51.05	10.32	47.79	97.34
		MSE	136.73	97.26	10.32	77.46	127.04
		Bias	-2.65	-4.20	0.04	3.42	1.95
	$\theta_0 = 0$	Var	34.76	13.85	10.40	15.05	28.90
		MSE	41.76	31.46	10.40	26.76	32.70
$b_0 = 4.5$		Bias	-4.18	-5.19	0.02	4.25	3.76
$b_1 = 3.5$	$\theta_0 = 0.5$	Var	66.39	24.66	10.32	24.88	55.59
$\sigma_e = 1$		MSE	83.86	51.62	10.32	42.92	69.69
		Bias	-4.73	-6.06	0.00	4.91	4.40
	$\theta_0 = 1$	Var	86.47	36.52	10.39	37.13	77.11
		MSE	108.80	73.27	10.39	61.21	96.47
		Bias	-2.73	-4.42	-0.07	3.72	2.52
	$\theta_0 = 0$	Var	26.57	10.74	10.01	10.44	22.37
		MSE	34.00	30.30	10.02	24.26	28.70
$b_0 = 2.5$		Bias	-3.60	-4.91	-0.07	4.17	3.26
$b_1 = 2.5$	$\theta_0 = 0.5$	Var	48.62	17.48	10.07	16.90	38.65
$\sigma_e = 0.5$		MSE	61.54	41.61	10.08	34.27	49.28
		Bias	-4.10	-5.28	-0.08	4.45	3.79
	$\theta_0 = 1$	Var	57.91	21.34	10.01	21.46	47.42
		MSE	74.68	49.24	10.02	41.28	61.78

Table 4: Bias($\widehat{F}_{\widetilde{e}}(t)$) ($\times 10^2$), Var($\widehat{F}_{\widetilde{e}}(t)$) ($\times 10^4$) and $MSE(\widehat{F}_{\widetilde{e}}(t))$ ($\times 10^4$) for different models, values of t and $n = 200$, when the distribution of e is a mixture of two normal distributions ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights.

Model	θ_0	$n = 100$	$n = 200$
$b_0 = 6.5$	$\theta_0 = 0$	35.86	17.80
$b_1 = 5$	$\theta_0 = 0.5$	80.74	38.92
$\sigma_e = 1.5$	$\theta_0 = 1$	105.91	51.24
$b_0 = 4.5$	$\theta_0 = 0$	37.23	17.19
$b_1 = 3.5$	$\theta_0 = 0.5$	66.28	30.42
$\sigma_e = 1$	$\theta_0 = 1$	83.10	40.66
$b_0 = 2.5$	$\theta_0 = 0$	43.50	14.87
$b_1 = 2.5$	$\theta_0 = 0.5$	65.54	23.19
$\sigma_e = 0.5$	$\theta_0 = 1$	81.86	27.40

Table 5: $IMSE(\widehat{F}_{\widehat{\varepsilon}}(t))$ ($\times 10^4$) for different models, values of t and sample sizes, when the distribution of e is a mixture of two normal distributions ($N(-1.5, 0.25)$, $N(1.5, 0.25)$) with equal weights.

6 Proofs

6.1 Auxiliary results

This section states a number of results concerning the estimators $\widehat{m}_{\widehat{\theta}}(x)$, $\widehat{m}_{\theta_o}(x)$ and $\Lambda_{\widehat{\theta}}(Y)$, which are needed for proving Theorem **??**. These results are of independent interest and their proofs can be found in Appendix A.

Proposition 6.1. *Assume* (A1)-(A9). *Then,*

$$\sup_{x\in\mathcal{X}_0} |\widehat{m}_{\widehat{\theta}}(x) - m_{\theta_o}(x)| = O_{\mathbb{P}}((nh)^{-1/2}(\log h^{-1})^{1/2}).$$

Proposition 6.2. *Under* (A1)-(A9), *we have*

$$\sup_{x\in\mathcal{X}_0} |\widehat{m}'_{\widehat{\theta}}(x) - m'_{\theta_o}(x)| = O_{\mathbb{P}}((nh^3)^{-1/2}(\log h^{-1})^{1/2}).$$

Proposition 6.3. *Assume* (A1)-(A9). *Then, for all* $\delta \in (0, 1)$,

$$\sup_{x,x'\in\mathcal{X}_0} \frac{|\widehat{m}'_{\widehat{\theta}}(x) - m'_{\theta_o}(x) - \widehat{m}'_{\widehat{\theta}}(x') + m'_{\theta_o}(x')|}{|x - x'|^{\delta}} = O_{\mathbb{P}}((nh^{3+2\delta})^{-1/2}(\log h^{-1})^{1/2}).$$

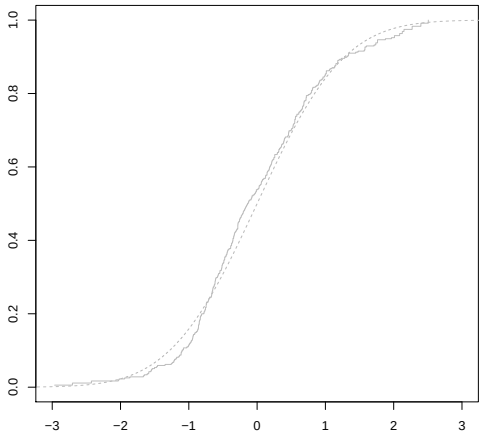

Figure 2: Error distribution for the Box-Cox model linking the scattering angle at which the polarization of sunlight vanishes to the cube root of particulate concentration. The solid line corresponds to our empirical estimator and the dashed line to the standard normal distribution.

Proposition 6.4. *Let* $\mathrm{Var}_n(\cdot)$ *be the conditional variance given* $\mathscr{X}_n$ *and assume that* (A1)-(A9) *hold. Then,*

$$\mathrm{Var}_n\left[\mathbb{1}\left(\Lambda_{\widehat{\theta}}(Y) \leq t + \widehat{m}_{\widehat{\theta}}(X)\right) - \mathbb{1}\left(\Lambda_{\theta_o}(Y) \leq t + m_{\theta_o}(X)\right)\right] = o_{\mathbb{P}}(1).$$

Proposition 6.5. *Assume* (A1)-(A9)*. Then,*

$$\int (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)) dF_X(x) = n^{-1}\sum_{i=1}^{n}(\Lambda_{\theta_o}(Y_i) - m_{\theta_o}(X_i)) + \frac{h^{q_1}}{q_1!}\mu(q_1, K_1)\mathbb{E}\left[m_{\theta_o}^{(q_1)}(X)\right] + o_{\mathbb{P}}(h^{q_1}),$$

where $m_{\theta_o}^{(q)}(x)$ *denotes the* $q-$*th derivative of* $m_{\theta_o}(x)$ *with respect to* x*.*

Proposition 6.6. *Assume* (A1)-(A9)*. Then,*

$$\begin{aligned}
&\mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \leq t | \mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \leq t\right) \\
&= \quad n^{-1}\sum_{i=1}^{n}\varepsilon_i f_\varepsilon(t) + (\widehat{\theta} - \theta_o)^t h(t) + h^{q_1}\frac{f_\varepsilon(t)}{q_1!}\mu(q_1, K_1)\mathbb{E}\left[m_{\theta_o}^{(q_1)}(X)\right] + R_n(t),
\end{aligned}$$

where $\sup\{|R_n(t)| : t \in \mathbb{R}\} = o(h^{q_1}) + o_{\mathbb{P}}(n^{-1/2})$*.*

The proofs of these propositions are given in Appendix A.

6.2 Proofs of the main results

This section contains the proofs of Theorem **??** and Corollary **??**. Some technical results needed in the proof of Theorem **??** are deferred to Appendices A and B.

Proof of Theorem ??

The result of the theorem directly follows from Lemma **??** in Appendix B and Proposition **??**. Indeed, using the latter results and the notations in the statement of the theorem, we have

$$\begin{aligned}
\widehat{F}_{\widehat{\varepsilon}}(t) - F_{\varepsilon}(t) &= n^{-1}\sum_{i=1}^{n}\left\{\mathbb{1}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \leq t\right) - F_{\varepsilon}(t)\right\} \\
&\quad + \mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \leq t | \mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \leq t\right) + o_{\mathbb{P}}(n^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\left\{\mathbb{1}\left(\varepsilon_i \leq t\right) - F_{\varepsilon}(t)\right\} \\
&\quad + n^{-1}\sum_{i=1}^{n}\varepsilon_i f_{\varepsilon}(t) + (\widehat{\theta} - \theta_o)^t h(t) + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}$$

where the last term $o_{\mathbb{P}}(n^{-1/2})$ is uniform in t. □

Proof of Corollary ??

To show the weak convergence of the process $\widehat{Z}_n(t)$ $(-\infty < t < +\infty)$, we make use of the techniques developed in Van der Vaart and Wellner (1996), involving the theory of bracketing numbers. In particular, we will show that (see Theorem 2.5.6 in that book)

$$\int_0^{\infty} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))}d\epsilon < \infty, \tag{6.1}$$

where $N_{[]}$ is the bracketing number, P is the probability measure corresponding to the joint distribution of (X, Y), $L_2(P)$ is the L_2-norm, and

$$\mathcal{F} = \left\{\phi_{\theta_o}(t, X, Y) : -\infty < t < +\infty\right\}.$$

Proving this entails that the class $\mathcal{F}$ is Donsker and hence the weak convergence of the given process follows from pp. 81-82 in Van der Vaart and Wellner's book. The two last terms of $\phi_{\theta_o}(t, X, Y)$ are the product of a random factor that is independent of t and a deterministic function, while the term $\mathbb{1}[\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \leq t]$ is decreasing in $\Lambda_{\theta_o}(Y) - m_{\theta_o}(X)$. Hence, $O(\exp(K\epsilon^{-1}))$ brackets are needed for this term by Theorem 2.7.5 in the aforementioned book. This concludes the proof, since the integration in (**??**) can be restricted to the

interval $[0, 2M]$, if the functions in the class $\mathcal{F}$ are bounded by M (for $\epsilon > 2M$ we take $N_{[]}(\epsilon, \mathcal{F}, L_2(P)) = 1$).

□

Appendix A. Proof of the auxiliary results

This appendix presents the proof of the propositions stated in Section 5.

Proof of Proposition ??

Let $c_n = (nh)^{-1/2}(\log h^{-1})^{1/2}$ and write

$$\widehat{m}_{\widehat{\theta}}(x) - m_{\theta_o}(x) = (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)) + \left(\widehat{m}_{\widehat{\theta}}(x) - \widehat{m}_{\theta_o}(x)\right).$$

We need to show that each of the above terms is $O_{\mathbb{P}}(c_n)$ uniformly in $x \in \mathcal{X}_0$. The term $\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)$ is treated by Lemma **??** in Appendix B. Consider $\widehat{m}_{\widehat{\theta}}(x) - \widehat{m}_{\theta_o}(x)$. Since $\widehat{\theta} - \theta_o = O_{\mathbb{P}}(n^{-1/2})$ by Theorem 4.1 in Linton, Sperlich and Van Keilegom (2008), a Taylor expansion applied to the function $\theta \to \widehat{m}_\theta(x)$, yields (to simplify notations, we assume here that $p = \dim(\theta) = 1$)

$$\begin{aligned}
\widehat{m}_{\widehat{\theta}}(x) - \widehat{m}_{\theta_o}(x) &= (\widehat{\theta} - \theta_o)\dot{\widehat{m}}_{\theta_o}(x) + \frac{1}{2}(\widehat{\theta} - \theta_o)^2 \ddot{\widehat{m}}_{\theta^*}(x) \\
&= O_{\mathbb{P}}(n^{-1/2})(nh\widehat{f}_X(x))^{-1} \sum_{i=1}^{n} \dot{\Lambda}_{\theta_o}(Y_i) K_1\left(\frac{X_i - x}{h}\right) \\
&\quad + O_{\mathbb{P}}(n^{-1})(nh\widehat{f}_X(x))^{-1} \sum_{i=1}^{n} \ddot{\Lambda}_{\theta^*}(Y_i) K_1\left(\frac{X_i - x}{h}\right), \qquad \text{(A.2)}
\end{aligned}$$

where θ^* is an intermediate value between θ_o and $\widehat{\theta}$, and where $\widehat{f}_X(x) = (nh)^{-1}\sum_{j=1}^{n} K_1(\frac{X_j - x}{h})$. Moreover, by Lemma 2 (in Appendix B), (A7)(i) and the Markov inequality, it can be shown that

$$(nh\widehat{f}_X(x))^{-1} \sum_{i=1}^{n} \dot{\Lambda}_{\theta_o}(Y_i) K_1\left(\frac{X_i - x}{h}\right) = O_{\mathbb{P}}(1), \quad (nh\widehat{f}_X(x))^{-1} \sum_{i=1}^{n} \ddot{\Lambda}_{\theta^*}(Y_i) K_1\left(\frac{X_i - x}{h}\right) = O_{\mathbb{P}}(h^{-1}),$$

uniformly in $x \in \mathcal{X}_0$. Substituting these orders in (**??**), gives

$$\widehat{m}_{\widehat{\theta}}(x) - \widehat{m}_{\theta_o}(x) = O_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(c_n),$$

uniformly in $x \in \mathcal{X}_0$ under (A2). This completes the proof of the proposition. □

Proof of Proposition ??

Let $c'_n = (\log h^{-1})^{1/2}(nh^3)^{-1/2}$ and write

$$\widehat{m}'_{\widehat{\theta}}(x) - m'_{\theta_o}(x) = (\widehat{m}'_{\theta_o}(x) - m'_{\theta_o}(x)) + (\widehat{m}'_{\widehat{\theta}}(x) - \widehat{m}'_{\theta_o}(x)). \qquad \text{(A.3)}$$

We need to show that each of the above terms is $O_{\mathbb{P}}(c'_n)$ uniformly in $x \in \mathcal{X}_0$. Consider the first term of (**??**) and note that $\mathbb{E}[\Lambda^4_{\theta_o}(Y)|X = x] \leq C\left(|m_{\theta_o}(x)|^4 + \mathbb{E}[\varepsilon^4]\right)$, for some $C > 0$. Since $\mathbb{E}[\varepsilon^4] < \infty$, the compactness of $\mathcal{X}_0$ and the continuity of m_{θ_o} ensure that $\mathbb{E}[\Lambda^4_{\theta_o}(Y)|X = x] < \infty$ uniformly in $x \in \mathcal{X}_0$. Then using arguments similar to Theorem 2 in Einmahl and Mason (2005) and Lemma 2 in Appendix B (extended to derivatives with respect to x) leads to $\sup_x |\widehat{m}'_{\theta_o}(x) - m'_{\theta_o}(x)| = O_{\mathbb{P}}(c'_n)$. For the second term of (**??**), we have similarly to the proof of Proposition **??** (for some θ^* between θ_o and $\widehat{\theta}$ and $p = 1$ to simplify notations)

$$\begin{aligned}
\widehat{m}'_{\widehat{\theta}}(x) - \widehat{m}'_{\theta_o}(x) &= (\widehat{\theta} - \theta_o)\dot{\widehat{m}}'_{\theta_o}(x) + (\widehat{\theta} - \theta_o)^2 \ddot{\widehat{m}}'_{\theta^*}(x) \\
&= (\widehat{\theta} - \theta_o)\frac{d}{dx}\left[\frac{\sum_{i=1}^n \dot{\Lambda}_{\theta_o}(Y_i)K_1\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K_1\left(\frac{X_i - x}{h}\right)}\right] \\
&\quad + \frac{(\widehat{\theta} - \theta_o)^2}{2}\frac{\widehat{f}_X(x)\frac{\partial}{\partial x}R(\ddot{\Lambda}_{\theta^*}, x) - R(\ddot{\Lambda}_{\theta^*}, x)\widehat{f}'_X(x)}{\widehat{f}^2_X(x)},
\end{aligned} \tag{A.4}$$

where $R(\Lambda, x) = \frac{1}{nh}\sum_{i=1}^n \Lambda(Y_i)K_1\left(\frac{X_i - x}{h}\right)$. Since $\widehat{\theta} - \theta_o = O_{\mathbb{P}}(n^{-1/2})$ by Theorem 4.1 of Linton, Sperlich and Van Keilegom (2008), the first term on the right hand side of the above expression is $O_{\mathbb{P}}(n^{-1/2})$ using the same arguments as above (Einmahl and Mason (2005) and Lemma 2 in Appendix B) applied to the data $\dot{\Lambda}_{\theta_o}(Y_i)$, $i = 1, \ldots, n$, while the second term is treated with assumptions (A3)(ii), (A7)(i) and the Markov inequality. This finishes the proof. □

Proof of Proposition ??

Let $\widetilde{c}_n = (\log h^{-1})^{1/2}(nh^{3+2\delta})^{-1/2}$, $d_n(x) = \widehat{m}_{\widehat{\theta}}(x) - m_{\theta_o}(x)$ and define $\beta_n(x, x') = |x - x'|^{-\delta}|d'_n(x) - d'_n(x')|$. We need to show that $\sup_{x,x'} |\beta_n(x, x')| = O_{\mathbb{P}}(\widetilde{c}_n)$. Note that by Proposition **??** the result is straightforward when $|x - x'| \geq Ch$, for some $C > 0$. Let us now consider x and x' such that $|x - x'| \leq Ch$. Then a Taylor expansion applied to $|d'_n(x) - d'_n(x')|$ gives

$$\begin{aligned}
|\beta_n(x, x')|\mathbb{1}\left(|x - x'| \leq Ch\right) &\leq \mathbb{1}\left(|x - x'| \leq Ch\right)|x - x'|^{1-\delta}\sup_x |d''_n(x)| \\
&\leq (Ch)^{1-\delta}\sup_x |d''_n(x)|,
\end{aligned}$$

so that the result of the proposition holds if $\sup_x |d''_n(x)| = O_{\mathbb{P}}((\log h^{-1})^{1/2}(nh^5)^{-1/2})$. For this, arguments similar to Einmahl and Mason (2005) and Lemma 2 in Appendix B (used in the same way as in Proposition **??**) enable to show that $\widehat{m}''_{\theta_o}(x) - m''_{\theta_o}(x) = O_{\mathbb{P}}((\log h^{-1})^{1/2}(nh^5)^{-1/2})$ uniformly in x. Moreover, in a completely similar way as done for (**??**) in the proof of Proposition **??**, it can be shown that $\widehat{m}''_{\widehat{\theta}}(x) - \widehat{m}''_{\theta_o}(x) = O_{\mathbb{P}}((\log h^{-1})^{1/2}(nh^5)^{-1/2})$ uniformly in x. This finishes the proof of the proposition. □

Proof of Proposition ??

Write

$$\begin{aligned}
&\mathrm{Var}_n\left[\mathbb{1}\left\{\Lambda_{\widehat{\theta}}(Y) \le t + \widehat{m}_{\widehat{\theta}}(X)\right\} - \mathbb{1}\left\{\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X)\right\}\right] \\
&\quad \le \quad 2\mathrm{Var}_n\left[\mathbb{1}\left\{\Lambda_{\widehat{\theta}}(Y) \le t + m_{\theta_o}(X) + d_n(X)\right\} - \mathbb{1}\left\{\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X) + d_n(X)\right\}\right] \\
&\qquad + 2\mathrm{Var}_n\left[\mathbb{1}\left\{\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X) + d_n(X)\right\} - \mathbb{1}\left\{\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X)\right\}\right]. \qquad \text{(A.5)}
\end{aligned}$$

We will show that each of the above terms is $o_{\mathbb{P}}(1)$ as $n \to \infty$. For the first term of (**??**), let $A_{n\widehat{\theta}}(x) = t + m_{\theta_o}(x) + d_n(x)$, $\Phi_n(\theta, x, y) = \mathbb{P}(Y \le V_\theta(y)|x, \mathscr{X}_n)$, $V_\theta(y) = \Lambda_\theta^{-1}(y)$ for all $\theta \in \Theta$ and write

$$\begin{aligned}
&\mathrm{Var}_n\left[\mathbb{1}\left\{\Lambda_{\widehat{\theta}}(Y) \le t + m_{\theta_o}(X) + d_n(X)\right\} - \mathbb{1}\left\{\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X) + d_n(X)\right\}\right] \\
&\quad \le \quad \mathbb{E}\left[\left(\mathbb{1}\left\{\Lambda_{\widehat{\theta}}(Y) \le A_{n\widehat{\theta}}(X)\right\} - \mathbb{1}\left\{\Lambda_{\theta_o}(Y) \le A_{n\widehat{\theta}}(X)\right\}\right)^2 |\mathscr{X}_n\right] \\
&\quad = \quad \int |F_{Y|X}(V_{\widehat{\theta}}(A_{n\widehat{\theta}}(x))|x, \mathscr{X}_n) - F_{Y|X}(V_{\theta_o}(A_{n\widehat{\theta}}(x))|x, \mathscr{X}_n)| dF_X(x) \\
&\quad = \quad (\widehat{\theta} - \theta_o)^t \int \left| \frac{\partial}{\partial \theta} F_{Y|X}(V_\theta(A_{n\widehat{\theta}}(x))|x, \mathscr{X}_n)\big|_{\theta=\theta^*} \right| dF_X(x),
\end{aligned}$$

for some θ^* between θ_o and $\widehat{\theta}$ and where $\frac{\partial}{\partial\theta} F_{Y|X}(V_\theta(\cdot)|x, \mathscr{X}_n)|_{\theta=\theta^*}$ denotes the vector of partial derivatives of $F_{Y|X}(V_\theta(\cdot)|x, \mathscr{X}_n)$ with respect to θ calculated at the point $\theta = \theta^*$. This term is thus $o_{\mathbb{P}}(1)$ by (A6) and the fact that $\widehat{\theta} - \theta_o = O_{\mathbb{P}}(n^{-1/2})$.

Consider now the second term of (**??**).

$$\begin{aligned}
&\mathrm{Var}_n\left[\mathbb{1}\left(\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X) + d_n(X)\right) - \mathbb{1}\left(\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X)\right)\right] \\
&\quad \le \quad \mathbb{E}\left[\left\{\mathbb{1}\left(Y \le V_{\theta_o}(t + m_{\theta_o}(X) + d_n(X))\right) - \mathbb{1}\left(Y \le V_{\theta_o}(t + m_{\theta_o}(X))\right)\right\}^2 |\mathscr{X}_n\right] \\
&\quad = \quad \int |F_{Y|X}(V_{\theta_o}(t + m_{\theta_o}(x) + d_n(x))|x, \mathscr{X}_n) - F_{Y|X}(V_{\theta_o}(t + m_{\theta_o}(x))|x)| dF_X(x) \\
&\quad \le \quad K \sup_x |d_n(x)| \sup_{\theta,x,y} \left| \frac{\partial}{\partial y} F_{Y|X}\left(V_\theta(y)|x\right) \right|,
\end{aligned}$$

for some $K > 0$. This term is $o_{\mathbb{P}}(1)$, since $\sup_x |d_n(x)| = o_{\mathbb{P}}(1)$ uniformly in x. This finishes the proof. □

Proof of Proposition ??

Let $c_n = (nh)^{-1/2}(\log h^{-1})^{1/2}$ and note that

$$\begin{aligned}
&\int (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)) dF_X(x) \\
&\quad = \quad \int \frac{\widehat{f}_X(x)}{f_X(x)} (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)) dF_X(x) + \int \left(\frac{f_X(x) - \widehat{f}_X(x)}{f_X(x)} \right) (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)) dF_X(x) \\
&\quad = \quad A_n + B_n, \qquad \text{(A.6)}
\end{aligned}$$

where $\widehat{f}_X(x) = (nh)^{-1}\sum_{j=1}^n K_1(\frac{X_j - x}{h})$. For the first term above, write

$$\begin{aligned} A_n &= \int \frac{\widehat{f}_X(x)}{f_X(x)}(\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x))dF_X(x) \\ &= (nh)^{-1}\sum_{i=1}^n \int (\Lambda_{\theta_o}(Y_i) - m_{\theta_o}(x))\, K_1\left(\frac{X_i - x}{h}\right)\frac{dF_X(x)}{f_X(x)} \\ &= (nh)^{-1}\sum_{i=1}^n \int (\Lambda_{\theta_o}(Y_i) - m_{\theta_o}(X_i))K_1\left(\frac{X_i - x}{h}\right)dx \\ &\quad + (nh)^{-1}\sum_{i=1}^n \int (m_{\theta_o}(X_i) - m_{\theta_o}(x))K_1\left(\frac{X_i - x}{h}\right)dx \\ &= A_{1n} + A_{2n}. \end{aligned} \tag{A.7}$$

Next,

$$\begin{aligned} A_{1n} &= (nh)^{-1}\sum_{i=1}^n \int (\Lambda_{\theta_o}(Y_i) - m_{\theta_o}(X_i))K_1\left(\frac{X_i - x}{h}\right)dx \\ &= n^{-1}\sum_{i=1}^n (\Lambda_{\theta_o}(Y_i) - m_{\theta_o}(X_i)). \end{aligned} \tag{A.8}$$

For the second term of (**??**), a Taylor expansion applied to $m_{\theta_o}(\cdot)$ yields

$$\begin{aligned} A_{2n} &= (nh)^{-1}\sum_{i=1}^n \int (m_{\theta_o}(X_i) - m_{\theta_o}(x))K_1\left(\frac{X_i - x}{h}\right)dx \\ &= n^{-1}\sum_{i=1}^n \int (m_{\theta_o}(X_i) - m_{\theta_o}(X_i - vh))K_1(v)dv \\ &= \frac{h^{q_1}}{q_1!}n^{-1}\sum_{i=1}^n m_{\theta_o}^{(q_1)}(X_i)\int v^{q_1}K_1(v)dv + o_{\mathbb{P}}(h^{q_1}). \end{aligned} \tag{A.9}$$

Hence by (**??**), (**??**), (**??**) and (**??**), the result of the proposition holds since $B_n = o_{\mathbb{P}}(h^{q_1})$ by assumption (A2). □

Proof of Proposition ??

Let $c_n = (nh)^{-1/2}(\log h^{-1})^{1/2}$ and write

$$\begin{aligned} &\mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \le t|\mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \le t\right) \\ &= \left[\mathbb{P}\left(\Lambda_{\theta_o}(Y) - \widehat{m}_{\theta_o}(X) \le t|\mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \le t\right)\right] \\ &\quad + \left[\mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \le t|\mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - \widehat{m}_{\theta_o}(X) \le t|\mathscr{X}_n\right)\right]. \end{aligned} \tag{A.10}$$

Consider the first term above. By Lemma **??** in Appendix B, we have $\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x) = O_{\mathbb{P}}(c_n)$ uniformly in x. Then, applying a Taylor expansion to $F_{Y|X}(V_{\theta_o}(\cdot)|x)$ and using assumption (A6),

$$\begin{aligned} F_{Y|X}\left(V_{\theta_o}(t+\widehat{m}_{\theta_o}(x))|x,\mathscr{X}_n\right) - F_{Y|X}\left(V_{\theta_o}(t+m_{\theta_o}(x))|x\right) &= (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x))\frac{\partial}{\partial t}F_{Y|X}(V_{\theta_o}(t+m_{\theta_o}(x))|x) \\ &\quad + O_{\mathbb{P}}(c_n^2), \end{aligned}$$

where the term $O_{\mathbb{P}}(c_n^2)$ is uniform in t and x. Therefore, since $f_\varepsilon(t) = \frac{\partial}{\partial t}F_{Y|X}(V_{\theta_o}(t+m_{\theta_o}(x))|x)$ for all x and $c_n^2 = o(h^{q_1})$,

$$\begin{aligned} &\mathbb{P}\left(\Lambda_{\theta_o}(Y) - \widehat{m}_{\theta_o}(X) \le t|\mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \le t\right) \\ &= \int [F_{Y|X}\left(V_{\theta_o}(t+\widehat{m}_{\theta_o}(x))|x,\mathscr{X}_n\right) - F_{Y|X}\left(V_{\theta_o}(t+m_{\theta_o}(x))|x\right)]dF_X(x) \\ &= \int (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x))\frac{\partial}{\partial t}F_{Y|X}(V_{\theta_o}(t+m_{\theta_o}(x))|x)dF_X(x) + O_{\mathbb{P}}(c_n^2) \\ &= f_\varepsilon(t)\int (\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x))dF_X(x) + O_{\mathbb{P}}(c_n^2) \\ &= f_\varepsilon(t)n^{-1}\sum_{i=1}^n \varepsilon_i + \frac{h^{q_1}}{q_1!}f_\varepsilon(t)\mu(q_1,K_1)\mathbb{E}[m_{\theta_o}^{(q_1)}(X)] + o_{\mathbb{P}}(h^{q_1}), \end{aligned} \tag{A.11}$$

using Proposition **??** and where $o_{\mathbb{P}}(h^{q_1})$ is uniform in t. For the second term of (**??**), let $\Phi_t(\theta,x,y,\mathscr{X}_n) = F_{Y|X}(V_\theta(t+y)|x,\mathscr{X}_n)$. Then, applying a Taylor expansion to the function $\theta \to \Phi_t(\theta,x,\widehat{m}_\theta(x),\mathscr{X}_n)$ and using (A6) and (A7)(i), we have

$$\begin{aligned} &\mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \le t|\mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - \widehat{m}_{\theta_o}(X) \le t|\mathscr{X}_n\right) \\ &= \int [\Phi_t(\widehat{\theta},x,\widehat{m}_{\widehat{\theta}}(x),\mathscr{X}_n) - \Phi_t(\theta_o,x,\widehat{m}_{\theta_o}(x)\mathscr{X}_n)]dF_X(x) \\ &= (\widehat{\theta}-\theta_o)^t\int \frac{d}{d\theta}\,\Phi_t(\theta,x,\widehat{m}_\theta(x),\mathscr{X}_n)|_{\theta=\theta_o}\,dF_X(x) + o_{\mathbb{P}}(n^{-1/2}) \\ &= (\widehat{\theta}-\theta_o)^t\int \frac{d}{d\theta}\,[\Phi_t(\theta,x,\widehat{m}_\theta(x),\mathscr{X}_n) - \Phi_t(\theta,x,m_\theta(x))]|_{\theta=\theta_o}\,dF_X(x) \\ &\quad + (\widehat{\theta}-\theta_o)^t\int \frac{d}{d\theta}\,\Phi_t(\theta,x,m_\theta(x))|_{\theta=\theta_o}\,dF_X(x) + o_{\mathbb{P}}(n^{-1/2}) \\ &= A_n + B_n + o_{\mathbb{P}}(n^{-1/2}), \end{aligned}$$

where $o_{\mathbb{P}}(n^{-1/2})$ is uniform in t. Using the uniform consistency of $\widehat{m}_{\theta_o}(x)$ and $\dot{\widehat{m}}_{\theta_o}(x)$ stated in Lemma **??** (Appendix B) and (A6),

$$A_n = (\widehat{\theta}-\theta_o)^t\int \frac{d}{d\theta}\,[\Phi_t(\theta,x,\widehat{m}_\theta(x)|\mathscr{X}_n) - \Phi_t(\theta,x,m_\theta(x))]|_{\theta=\theta_o}\,dF_X(x) = o_{\mathbb{P}}(n^{-1/2}).$$

Therefore

$$\begin{aligned}
&\mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \leq t | \mathscr{X}_n\right) - \mathbb{P}\left(\Lambda_{\theta_o}(Y) - \widehat{m}_{\theta_o}(X) \leq t | \mathscr{X}_n\right) = B_n + o_{\mathbb{P}}(n^{-1/2}) \\
&= (\widehat{\theta} - \theta_o)^t \int \frac{d}{d\theta}\, \Phi_t(\theta, x, m_\theta(x))|_{\theta=\theta_o}\, dF_X(x) + o_{\mathbb{P}}(n^{-1/2}) \\
&= (\widehat{\theta} - \theta_o)^t \mathbb{E}\left[\frac{d}{d\theta}\, F_{Y|X}\left(V_\theta(t + m_\theta(X))|X\right)\big|_{\theta=\theta_o}\right] + o_{\mathbb{P}}(n^{-1/2}) \\
&= (\widehat{\theta} - \theta_o)^t h(t) + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}$$

where the term $o_{\mathbb{P}}(n^{-1/2})$ is uniform in $t \in \mathbb{R}$. The result of the proposition now follows from the above equality, (**??**) and (**??**). □

Appendix B

We start this appendix with a technical result needed in the proof of Theorem **??**.

Lemma 1. *Assume* (A1)-(A9). *Then,*

$$\begin{aligned}
n^{-1} \sum_{i=1}^{n} \Big\{ &\mathbb{1}\left(\Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\theta}}(X_i) \leq t\right) - \mathbb{1}\left(\Lambda_{\theta_o}(Y_i) - m_{\theta_o}(X_i) \leq t\right) \\
&- \mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) \leq t | \mathscr{X}_n\right) + \mathbb{P}\left(\Lambda_{\theta_o}(Y) - m_{\theta_o}(X) \leq t\right) \Big\} = o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}$$

uniformly for $t \in \mathbb{R}$.

Proof

Note that $\Lambda_{\widehat{\theta}}(Y) - \widehat{m}_{\widehat{\theta}}(X) = \Lambda_{\widehat{\theta}}(Y) - m_{\theta_o}(X) - d_n(X)$, where $d_n(X) = \widehat{m}_{\widehat{\theta}}(X) - m_{\theta_o}(X)$. The proof of the lemma is based on results in Van der Vaart and Wellner (1996). Define

$$\begin{aligned}
\mathscr{F}_1 = \Big\{ &(x, y) \to \mathbb{1}\left(\Lambda_\theta(y) \leq t + m_{\theta_o}(x) + d(x)\right),\ \Lambda_\theta : \mathbb{R} \to \mathbb{R} \text{ strictly increasing}, \\
&\theta \in \Theta,\ t \in \mathbb{R} \text{ and } d \in C_1^{1+\delta}(\mathcal{X}_0) \Big\}.
\end{aligned}$$

We observe that by Propositions **??**, **??** and **??**, we have $\mathbb{P}\left(d_n \in C_1^{1+\delta}(\mathcal{X}_0)\right) \to 1$ as $n \to \infty$. In a first step, we will show that the class $\mathscr{F}_1$ is Donsker. From Theorem 2.5.6 in Van der Vaart and Wellner (1996), it follows that it suffices to show that

$$\int_0^\infty \sqrt{\log N_{[]}(\bar{\varepsilon}, \mathscr{F}_1, L_2(P))}\, d\bar{\varepsilon} < \infty, \tag{B.1}$$

where $N_{[]}$ is the bracketing number, P is the probability measure corresponding to the joint distribution of (Y, X), and $L_2(P)$ is the L_2-norm.

Embed Θ into a hypercube $[\theta_1^\ell, \theta_1^u] \times \cdots \times [\theta_p^\ell, \theta_p^u]$ of dimension p, and for each $j = 1, \ldots, p$, let $\theta_j^\ell = \theta_{0j} \leq \theta_{1j} \leq \ldots \leq \theta_{m_j j} = \theta_j^u$ partition the finite interval $[\theta_j^\ell, \theta_j^u]$ into $m_j = O(\bar{\varepsilon}^{-2})$ intervals of length $O(\bar{\varepsilon}^2)$. This results in a partition of Θ into at most $\prod_{j=1}^p m_j = O(\bar{\varepsilon}^{-2p})$ hypercubes, which we denote by R_i, $i = 1, \ldots, \prod_{j=1}^p m_j$. For each nonempty R_i, let $\Gamma_i^\ell(Y) = \min_{\theta \in R_i \cap \Theta} \Lambda_\theta(Y)$ and $\Gamma_i^u(Y) = \max_{\theta \in R_i \cap \Theta} \Lambda_\theta(Y)$.

For the class $C_1^{1+\delta}(\mathcal{X}_0)$, Corollary 2.7.2 in Van der Vaart and Wellner (1996) ensures that

$$\log r := \log N_{[]}\left(\bar{\varepsilon}^2, C_1^{1+\delta}(\mathcal{X}_0), \|\cdot\|_\infty\right) \leq K \bar{\varepsilon}^{-2/(1+\delta)},$$

for some $K > 0$.

Let $d_1^\ell \leq d_1^u, \ldots, d_r^\ell \leq d_r^u$ be the functions defining the r brackets for the class $C_1^{1+\delta}(\mathcal{X}_0)$. Then, for each $\theta \in \Theta$ and each $d \in C_1^{1+\delta}(\mathcal{X}_0)$, there exist i and j such that

$$\begin{aligned} & \mathbb{1}\left\{\Gamma_i^u(Y) \leq t + m_{\theta_o}(X) + d_j^\ell(X)\right\} \\ \leq \quad & \mathbb{1}\left\{\Lambda_\theta(Y) \leq t + m_{\theta_o}(X) + d(X)\right\} \\ \leq \quad & \mathbb{1}\left\{\Gamma_i^\ell(Y) \leq t + m_{\theta_o}(X) + d_j^u(X), \right\}. \end{aligned}$$

Define

$$p_{ij}^{u\ell}(t) = \mathbb{P}\left(\Gamma_i^u(Y) \leq t + m_{\theta_o}(X) + d_j^\ell(X)\right)$$

and let $t_{ijk}^{u\ell}$, $k = 1, \ldots, O(\bar{\varepsilon}^{-2})$, partition the line in segments having $p_{ij}^{u\ell}$-probability less than or equal to a fraction of $\bar{\varepsilon}^2$. Similarly, define

$$p_{ij}^{\ell u}(t) = \mathbb{P}\left(\Gamma_i^\ell(Y) \leq t + m_{\theta_o}(X) + d_j^u(X)\right)$$

and let $t_{ijk}^{\ell u}$, $k = 1, \ldots, O(\bar{\varepsilon}^{-2})$, partition the line in segments having $p_{ij}^{\ell u}$-probability less than or equal to a fraction of $\bar{\varepsilon}^2$. Let us now define the following brackets for t:

$$t_{ijk_1}^{u\ell} \leq t \leq t_{ijk_2}^{\ell u},$$

where $t_{ijk_1}^{u\ell}$ is the largest of the $t_{ijk}^{u\ell}$ with the property of being less than or equal to t, and $t_{ijk_2}^{\ell u}$ is the smallest of the $t_{ijk}^{\ell u}$ with the property of being larger than or equal to t. We will now show that the $\bar{\varepsilon}$-brackets for our function are given by

$$\begin{aligned} & \mathbb{1}\left\{\Gamma_i^u(Y) \leq t_{ijk_1}^{u\ell} + m_{\theta_o}(X) + d_j^\ell(X)\right\} \\ \leq \quad & \mathbb{1}\left\{\Gamma(Y) \leq t + m_{\theta_o}(X) + d(X)\right\} \\ \leq \quad & \mathbb{1}\left\{\Gamma_i^\ell(Y) \leq t_{ijk_2}^{\ell u} + m_{\theta_o}(X) + d_j^u(X)\right\}. \end{aligned}$$

To this end, let us calculate

$$\begin{aligned}
&\left\|\mathbb{1}\left\{\Gamma_i^\ell(Y) \le t_{ijk_2}^{\ell u} + m_{\theta_o}(X) + d_j^u(X)\right\} - \mathbb{1}\left\{\Gamma_i^u(Y) \le t_{ijk_1}^{u\ell} + m_{\theta_o}(X) + d_j^\ell(X)\right\}\right\|_2^2 \\
&= \mathbb{P}\left(\Gamma_i^\ell(Y) \le t_{ijk_2}^{\ell u} + m_{\theta_o}(X) + d_j^u(X)\right) - \mathbb{P}\left(\Gamma_i^u(Y) \le t_{ijk_1}^{u\ell} + m_{\theta_o}(X) + d_j^\ell(X)\right) \\
&= p_{ij}^{\ell u}(t) - p_{ij}^{u\ell}(t) + O(\bar{\varepsilon}^2),
\end{aligned}$$

where $\|\cdot\|_2 = \|\cdot\|_{P,2}$ is the $L_2(P)$-norm. Since $\Gamma_i^\ell(y)$ and $\Gamma_i^u(y)$, $i = 1, \ldots, \prod_{j=1}^p m_j$, are strictly increasing continuous functions of $y \in \mathbb{R}$, they have inverse functions $\Gamma_i^{\ell^{-1}}(\cdot)$ and $\Gamma_i^{u^{-1}}(\cdot)$. Moreover, it is easy to check that $\Gamma_i^{\ell^{-1}}(\cdot) = \max_{\theta \in R_i} V_\theta(\cdot)$ and $\Gamma_i^{u^{-1}}(\cdot) = \min_{\theta \in R_i} V_\theta(\cdot)$. Therefore,

$$\begin{aligned}
&p_{ij}^{\ell u}(t) - p_{ij}^{u\ell}(t) \\
&= \int \Big[\mathbb{P}\left\{\Gamma_i^\ell(Y) \le t + m_{\theta_o}(x) + d_j^u(x) | X = x\right\} - \mathbb{P}\left\{\Gamma_i^u(Y) \le t + m_{\theta_o}(x) + d_j^\ell(x) | X = x\right\}\Big] dF_X(x) \\
&= \int \Big[F_{Y|X}(\Gamma_i^{\ell^{-1}}(t + m_{\theta_o}(x) + d_j^u(x))|x) - F_{Y|X}(\Gamma_i^{u^{-1}}(t + m_{\theta_o}(x) + d_j^u(x))|x) \\
&\qquad + F_{Y|X}(\Gamma_i^{u^{-1}}(t + m_{\theta_o}(x) + d_j^u(x))|x) - F_{Y|X}(\Gamma_i^{u^{-1}}(t + m_{\theta_o}(x) + d_j^\ell(x))|x)\Big] dF_X(x) \\
&\le \int \left[\sum_{q=1}^p \sup_{\theta \in \Theta, y \in \mathbb{R}} \left|\frac{\partial F_{Y|X}(V_\theta(t + m_{\theta_o}(x) + y)|x)}{\partial \theta_q}\right| \bar{\varepsilon}^2 + \sup_{\theta \in \Theta, y \in \mathbb{R}} \left|\frac{\partial F_{Y|X}(V_\theta(t + m_{\theta_o}(x) + y)|x)}{\partial y}\right| \bar{\varepsilon}^2\right] dF_X(x) \\
&= O(\bar{\varepsilon}^2),
\end{aligned}$$

using assumption (A6). That leads to

$$\left\|\mathbb{1}\left\{\Gamma_i^\ell(Y) \le t_{ijk_2}^{\ell u} + m_{\theta_o}(X) + d_j^u(X)\right\} - \mathbb{1}\left\{\Gamma_i^u(Y) \le t_{ijk_1}^{u\ell} + m_{\theta_o}(X) + d_j^\ell(X)\right\}\right\|_2^2 = O(\bar{\varepsilon}^2).$$

Hence, for each $\bar{\varepsilon} > 0$, we need at most $O(\bar{\varepsilon}^{-2(p+1)} \exp(K\bar{\varepsilon}^{-2/(1+\delta)}))$ brackets (for some $K > 0$) to cover the class $\mathscr{F}_1$. However, for $\bar{\varepsilon} > 1$, one bracket suffices. So we have

$$\int_0^\infty \sqrt{\log N_{[]}(\bar{\varepsilon}, \mathscr{F}_1, L_2(P))} d\bar{\varepsilon} < \infty,$$

which gives (**??**). This shows that the class $\mathscr{F}_1$ is Donsker, and hence by straightforward calculations,

$$\begin{aligned}
\mathscr{F} = \Big\{&(x, y) \to \mathbb{1}\left(\Lambda_\theta(y) \le t + m_{\theta_o}(x) + d(x)\right) - \mathbb{1}\left(\Lambda_{\theta_o}(y) \le t + m_{\theta_o}(x)\right) \\
&- \mathbb{P}\left(\Lambda_\theta(Y) \le t + m_{\theta_o}(X) + d(X)\right) + \mathbb{P}\left(\Lambda_{\theta_o}(Y) \le t + m_{\theta_o}(X)\right),\ \theta \in \Theta,\ t \in \mathbb{R},\ d \in C_\delta^{1+\delta}(\mathcal{X}_0)\Big\}
\end{aligned}$$

is a Donsker class as well.

Next, observe that for $d_n(X) = \widehat{m}_{\widehat{\theta}}(X) - m_{\theta_o}(X)$, Proposition **??** ensures that

$$\begin{aligned}
\mathrm{Var}_n\Big[&\mathbb{1}\left(\Lambda_{\widehat{\theta}}(Y) \leq t + m_{\theta_o}(X) + d_n(X)\right) - \mathbb{1}\left(\Lambda_{\theta_o}(Y) \leq t + m_{\theta_o}(X)\right) \\
&-\mathbb{P}\left(\Lambda_{\widehat{\theta}}(Y) \leq t + m_{\theta_o}(X) + d_n(X) | \mathscr{X}_n\right) + \mathbb{P}\left(\Lambda_{\theta_o}(Y) \leq t + m_{\theta_o}(X)\right)\Big] \\
= \quad & \mathrm{Var}_n\left[\mathbb{1}\left(\Lambda_{\widehat{\theta}}(Y) \leq t + m_{\theta_o}(X) + d_n(X)\right) - \mathbb{1}\left(\Lambda_{\theta_o}(Y) \leq t + m_{\theta_o}(X)\right)\right] = o_{\mathbb{P}}(1)
\end{aligned}$$

as $n \to \infty$. Since the class $\mathscr{F}$ is Donsker, it then follows from Corollary 2.3.12 in Van der Vaart and Wellner (1996) that

$$\lim_{\alpha \downarrow 0} \limsup_{n \to \infty} \mathbb{P}\left(\sup_{f \in \mathscr{F}, \mathrm{Var}(f) < \alpha} n^{-1/2} \left| \sum_{i=1}^{n} f(X_i) \right| > \bar{\varepsilon} \right) = 0,$$

for each $\bar{\varepsilon} > 0$. Hence by restricting the supremum inside the above probability to the elements in $\mathscr{F}$ corresponding to $d(X) = d_n(X)$ as defined above, the result of the lemma follows. □

Lemma 2. *Assume* (A1)-(A5) *and* (A7). *Then,*

$$\begin{aligned}
\sup_{x \in \mathcal{X}_0} |\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)| &= O_{\mathbb{P}}((nh)^{-1/2}(\log h^{-1})^{1/2}), \\
\sup_{x \in \mathcal{X}_0} |\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x)| &= O_{\mathbb{P}}((nh)^{-1/2}(\log h^{-1})^{1/2}).
\end{aligned}$$

Proof

We only give the proof for the uniform consistency of $\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x)$, the proof for $\widehat{m}_{\theta_o}(x) - m_{\theta_o}(x)$ being very similar. Let $c_n = (nh)^{-1/2}(\log h^{-1})^{1/2}$, and define

$$\dot{\widehat{r}}_{\theta_o}(x) = \frac{1}{nh} \sum_{j=1}^{n} \dot{\Lambda}_{\theta_o}(Y_j) K_1\left(\frac{X_j - x}{h}\right), \quad \dot{\overline{r}}_{\theta_o}(x) = \mathbb{E}[\dot{\widehat{r}}_{\theta_o}(x)], \quad \overline{f}_X(x) = \mathbb{E}[\widehat{f}_X(x)],$$

where $\widehat{f}_X(x) = (nh)^{-1} \sum_{j=1}^{n} K_1(\frac{X_j - x}{h})$. Then,

$$\sup_{x \in \mathcal{X}_0} |\dot{\widehat{m}}_{\theta_o}(x) - \dot{m}_{\theta_o}(x)| \leq \sup_{x \in \mathcal{X}_0} \left| \dot{\widehat{m}}_{\theta_o}(x) - \frac{\dot{\overline{r}}_{\theta_o}(x)}{\overline{f}_X(x)} \right| + \sup_{x \in \mathcal{X}_0} \frac{1}{\overline{f}_X(x)} \left| \dot{\overline{r}}_{\theta_o}(x) - \overline{f}_X(x) \dot{m}_{\theta_o}(x) \right|. \tag{B.2}$$

Since $\mathbb{E}[\dot{\Lambda}^4_{\theta_o}(Y) | X = x] < \infty$ uniformly in $x \in \mathcal{X}$ by assumption (A7), a similar proof as was given for Theorem 2 in Einmahl and Mason (2005) ensures that

$$\sup_{x \in \mathcal{X}_0} \left| \dot{\widehat{m}}_{\theta_o}(x) - \frac{\dot{\overline{r}}_{\theta_o}(x)}{\overline{f}_X(x)} \right| = O_{\mathbb{P}}\left(c_n\right).$$

Consider now the second term of (**??**). Since $\mathbb{E}[\dot{\varepsilon}(\theta_o)|X] = 0$, where $\dot{\varepsilon}(\theta_o) = \frac{d}{d\theta}(\Lambda_\theta(Y) - m_\theta(X))|_{\theta=\theta_o}$, we have

$$\begin{aligned}
\dot{\overline{r}}_{\theta_o}(x) &= h^{-1}\mathbb{E}\left[\left\{\dot{m}_{\theta_o}(X) + \dot{\varepsilon}(\theta_o)\right\} K_1\left(\frac{X-x}{h}\right)\right] \\
&= h^{-1}\mathbb{E}\left[\dot{m}_{\theta_o}(X) K_1\left(\frac{X-x}{h}\right)\right] \\
&= \int \dot{m}_{\theta_o}(x+hv) K_1(v) f_X(x+hv) dv,
\end{aligned}$$

from which it follows that

$$\dot{\overline{r}}_{\theta_o}(x) - \overline{f}_X(x)\dot{m}_{\theta_o}(x) = \int \left[\dot{m}_{\theta_o}(x+hv) - \dot{m}_{\theta_o}(x)\right] K_1(v) f_X(x+hv) dv.$$

Hence, Taylor expansions applied to $\dot{m}_{\theta_o}(\cdot)$ and $f_X(\cdot)$ yield

$$\sup_{x\in\mathcal{X}_0} \left|\dot{\overline{r}}_{\theta_o}(x) - \overline{f}_X(x)\dot{m}_{\theta_o}(x)\right| = O(h^{q_1}) = O\left(c_n\right),$$

since $nh^{2q_1+1}(\log h^{-1})^{-1} = O(1)$ by (A2). This proves that the second term of (**??**) is $O(c_n)$, since it can be shown that for h small enough $\bar{f}_X(\cdot)$ is bounded away from 0 and infinity uniformly on $\mathcal{X}$. □

References

[1] Akritas, M.G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.*, **28**, 549–567.

[2] Bellver, C. (1987). Influence of particulate pollution on the positions of neutral points in the sky in Seville (Spain). *Atmospheric Environment*, **21**, 699-702

[3] Bickel, P.J. and Doksum, K. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.,* **76**, 296–311.

[4] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley.

[5] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. - Ser. B*, **26**, 211–252.

[6] Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York.

[7] Chen, G., Lockhart, R.A. and Stephens, A. (2002). Box-Cox transformations in linear models: Large sample theory and tests of normality (with discussion). *Canad. J. Statist,* **30**, 177–234.

[8] Cheng, F. and Sun, S. (2008). A goodness-of-fit test of the errors in nonlinear autoregressive time series models. *Statist. Probab. Letters*, **78**, 50–59.

[9] Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, Summit.

[10] Dette, H., Neumeyer, N. and Van Keilegom, I. (2007). A new test for the parametric form of the variance function in nonparametric regression. *J. Royal Statist. Soc. - Series B*, **69**, 903–917.

[11] Dette, H., Pardo-Fernández, J.C. and Van Keilegom, I. (2009). Goodness-of-fit tests for multiplicative models with dependent data. *Scand. J. Statist.*, **36**, 782–799.

[12] Ding, Y. and Nan, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.*, **6**, 3032-3061.

[13] Einmahl, J. and Van Keilegom, I. (2008a). Tests for independence in nonparametric regression. *Statist. Sinica*, **18**, 601-616.

[14] Einmahl, J. and Van Keilegom, I. (2008b). Specification tests in nonparametric regression. *J. Econometrics*, **143**, 88-102.

[15] Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.,* **33**, 1380–1403.

[16] Fitzenberger, B., Wilke, R.A. and Zhang, X. (2010). Implementing Box-Cox quantile regression. *Econometric Rev.,* **29**, 158–181.

[17] Florens, J.-P., Simar, L. and Van Keilegom, I. (2014). Frontier estimation in nonparametric location-scale models. *J. Econometrics*, **178**, 456-470.

[18] González-Manteiga, W., Pardo-Fernández, J.C. and Van Keilegom, I. (2011). ROC curves in nonparametric location-scale regression models. *Scand. J. Statist.*, **38**, 169–184.

[19] Freeman, J. and Modarres, R. (2005). Efficiency of test for independence after Box-Cox transformation. *J. Multivar. Anal.*, **95**, 107–118.

[20] Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. Springer, New-York.

[21] Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory,* **24**, 726–748.

[22] Heuchenne, C. and Van Keilegom, I. (2010). Goodness-of-fit tests for the error distribution in nonparametric regression. *Comput. Statist. Data Anal.*, **54**, 1942–1951.

[23] Hlávka, Z., Husková, M. and Meintanis, S.G. (2011). Tests for independence in nonparametric heteroscedastic regression models. *J. Multiv. Anal.*, **102**, 816–827.

[24] Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Ann. Statist.*, **36**, 686–718.

[25] Manly, B.F. (1976). Exponential data transformation. *The Statistician*, **25**, 37–42.

[26] Müller, U.U., Schick, A. and Wefelmeyer, W. (2004). Estimating linear functionals of the error distribution in nonparametric regression. *J. Statist. Plann. Infer.,* **119**, 75–93.

[27] Müller, U.U., Schick, A. and Wefelmeyer, W. (2007). Estimating the error distribution function in semiparametric regression. *Statistics & Decisions*, **25**, 1–18.

[28] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, 141—142.

[29] Neumeyer, N. (2009a). Smooth residual bootstrap for empirical processes of nonparametric regression residuals. *Scand. J. Statist.*, **36**, 204–228.

[30] Neumeyer, N. (2009b). Testing independence in nonparametric regression. *J. Multiv. Anal.*, **100**, 1551–1566.

[31] Neumeyer, N. and Dette, H. (2007). Testing for symmetric error distribution in nonparametric regression models. *Statist. Sinica*, **17**, 775–795.

[32] Neumeyer, N., Pardo-Fernández, J.P. (2009). A simple test for comparing regression curves versus one-sided alternatives. *J. Statist. Plann. Infer.*, **139**, 4006–4016.

[33] Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multiv. Anal.,* **101**, 1067–1078.

[34] Pardo-Fernández, J.C., Van Keilegom, I. and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statist. Sinica*, **17**, 1115–1137.

[35] Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scand. J. Statist.*, **8**, 116–119.

[36] Sakia, R.M. (1992). The Box-Cox transformation technique: a review. *The Statistician,* **41**, 169–178.

[37] Shin, Y. (2008). Semiparametric estimation of the Box-Cox transformation model. *Econometrics J.,* **11**, 517–537.

[38] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, Chapman and Hall, London.

[39] Vanhems, A. and Van Keilegom, I. (2013). Semiparametric transformation model with endogeneity: a control function approach (submitted).

[40] Van Keilegom, I., González-Manteiga, W. and Sánchez Sellero, C. (2008). Goodness of fit tests in parametric regression based on the estimation of the error distribution. *TEST*, **17**, 401-415.

[41] Watson, G.S. (1964). Smooth regression analysis. *Sankhyā - Ser. A*, **26**, 359–372.

[42] Zellner, A. and Revankar, N.S. (1969). Generalized production functions. *Rev. Economic Studies,* **36**, 241–250.

[43] Zhang, C.M. (2003). Adaptative tests of regression functions via multiscale generalized likelihood ratio. *Canadian Journal of Statistics*, **31**, 151-171.